Received: 19 February 2020 | Accepted: 4 September 2020

DOI: 10.1111/rssc.12449

ROYAL STATISTICAL SOCIETY
Series C Applied Statistics

ORIGINAL ARTICLE

Functional ensemble survival tree: Dynamic prediction of Alzheimer's disease progression accommodating multiple time-varying covariates

Shu Jiang[1] | **Yijun Xie**[2] | **Graham A. Colditz**[1]

[1]Division of Public Health Sciences, Washington University School of Medicine in St. Louis, St. Louis, USA

[2]Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Canada

Correspondence
Shu Jiang, Division of Public Health Sciences, Washington University School of Medicine in St. Louis, St. Louis, USA.
Email: jiang.shu@wustl.edu

Funding information
Foundation for Barnes Jewish Hospital, Grant/Award Number: NIH P30 CA091842

Abstract

With the exponential growth in data collection, multiple time-varying biomarkers are commonly encountered in clinical studies, along with a rich set of baseline covariates. This paper is motivated by addressing a critical issue in the field of Alzheimer's disease (AD) in which we aim to predict the time for AD conversion in people with mild cognitive impairment to inform prevention and early treatment decisions. Conventional joint models of biomarker trajectory with time-to-event data rely heavily on model assumptions and may not be applicable when the number of covariates is large. This motivated us to consider a functional ensemble survival tree framework to characterize the joint effects of both functional and baseline covariates in predicting disease progression. The proposed framework incorporates multivariate functional principal component analysis to characterize the changing patterns of multiple time-varying neurocognitive biomarker trajectories and then nest these features within an ensemble survival tree in predicting the progression of AD. We provide a fast implementation of the algorithm that accommodates personalized dynamic prediction that can be updated as new observations are gathered to reflect the patient's latest prognosis. The algorithm is empirically shown to perform well in simulation studies and is illustrated through the analysis of data from

the Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu/). We provide implementation of our proposed method in an R package funest.

KEYWORDS

Time-varying covariates, Dynamic prediction, Functional principal component analysis, Random survival forest

1 | INTRODUCTION

Alzheimer's disease (AD) is one of the most prevalent diseases worldwide which leads to memory loss and dementia (LaFerla et al., 2007; Mattson, 2004; Rabin et al., 2019). Early detection is critical due to the lack of disease-modifying agents for patients diagnosed with AD. Mild cognitive impairment (MCI) is defined as the transition stage between the clinically normal and dementia state where it involves memory and language loss that is considered greater than expected age-related changes (Mattson, 2004). As a result, MCI patients are typically enrolled as the target population for early prognosis and evaluation of therapies (Ewers et al., 2012). There is considerable interest in identifying biomarkers or combinations of covariates, so that the likelihood of predicting the neurodegenerative pathology due to Alzheimer's disease for patients diagnosed with MCI can be greater. See Park et al. (2012); Ewers et al. (2012); Gomar et al. (2014) for example. Accurate prediction of disease progression to AD is thus important and critical to move the field forward (Kong et al., 2015; Risacher et al., 2009).

Tremendous amounts of data are being collected in the hopes of finding significant factors that may be associated with AD progression. In the dataset that motivated this work, the Alzheimer's Disease Neuroimaging Initiative (ADNI), the focus was on the collection of longitudinal assessments, magnetic resonance imaging and positron emission tomography imaging measures, as well as other biomarkers from blood and cerebrospinal fluid (Cuingnet et al., 2011). Of those covariates collected in the cohort, many are time-varying. For example, the cognitive change in preclinical AD is a series of cognitive tests which are measured at each patient visits. The potential for discovery would be much greater in incorporating all available patient-specific covariates in predicting the progression of AD. However, the challenges may arise from (a) high dimensionality of the baseline covariates; (b) presence of multivariate time-varying biomarkers; (c) non-linear and complex relationship between the covariates and the time-to-event outcome. A natural question then is how to best utilize this information to improve prediction performance to inform prevention and early treatment decisions.

Existing methods in the literature, such as the joint model, hinges on the pre-specified model assumptions for both the time-varying biomarker and the survival outcome (Rizopoulos, 2012). However, the nature of time-varying biomarkers may vary under different clinical settings, making it difficult to identify a suitable model. For illustration purposes, we present in Figure 1, the raw longitudinal trajectories for two of the longitudinal cognitive measures for 50 randomly selected MCI patients in ADNI. We can see that both the Mini Mental State Examination (MMSE, left) and Functional Activities Questionnaire (FAQ, right) trajectories have changing patterns over time and are highly variable within and between patients. In addressing this concern, nonparametric methods such as splines or kernel smoothing, have been adopted in the literature for prediction using the denoised smoothed values of the biomarker trajectories (Welsh et al., 2002; Wu & Chiang, 2000). More recently, functional approaches such as functional principal component analysis (FPCA), has become a

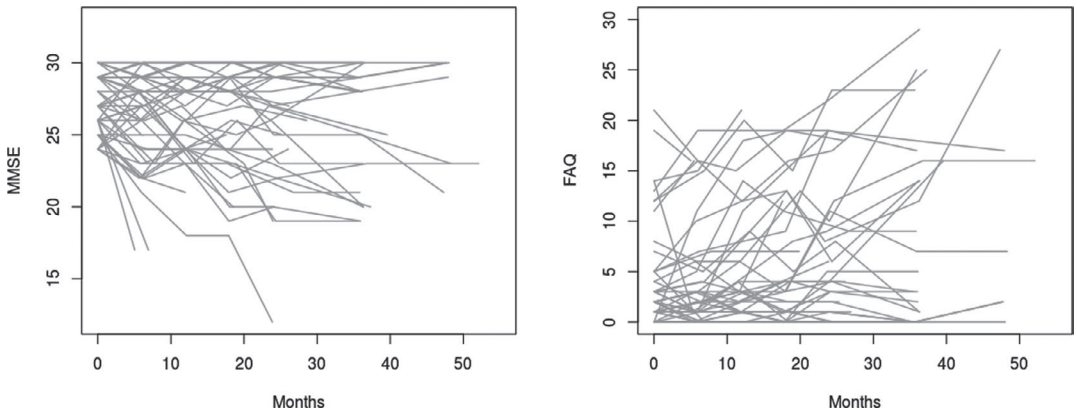

FIGURE 1 Longitudinal trajectories of Mini Mental State Examination (MMSE, left) and Functional Activities Questionnaire (FAQ, right) of 50 randomly selected Mild cognitive impairment patients in Alzheimer's Disease Neuroimaging Initiative

popular alternative for modelling time-varying predictors due to its ability to use extracted features in addition to the denoised smoothed values which will likely improve prediction (Ramsay & Silverman, 2004; Wang et al., 2016). Examples of utilizing functional data analysis in predicting the time-to-event outcome include Yan et al. (2017, 2018); Kong et al. (2018). However, all of the aforementioned methods focus on the dynamic prediction of time-to-event outcome with a single time-varying biomarker. As a result, Li and Luo (2019) recently proposed to use multiple longitudinal biomarkers in predicting the disease progression. However, their method was contingent on the proportional hazards model which may not be viable especially when the number of covariates is large. At the time of revision, we are aware of a recent article utilizing multiple time-varying biomarkers under a functional survival forest framework (Lin et al., 2020). However, their method did not consider the selection of an optimal estimator, which could lead to sub-optimal prediction performance.

In this article, we propose a unified strategy for dynamic prediction that does not depend on the model specification and can handle high-dimensional baseline and multivariate time-varying covariates in the presence of right censoring. The proposed approach is entirely data-driven and can be stated in terms of three main steps. (1) First, extract features from the multivariate time-varying covariates such that the changing patterns can be summarized by a set of basis functions and the associated individualized functional scores. (2) Then, construct candidate estimators based on the extracted features and observed data. (3) Last, apply cross-validation to select the optimal estimator among all candidates in step 2. Specifically, we adopt tree-based methods in this paper where the possible candidate estimators in step 2 are generated by repeated binary recursive partitions (Ishwaran et al., 2008, 2011). Tree-based methods facilitate a comprehensive modelling scheme and are appealing for their ability to handle high-dimensional covariates, facilitate complex and non-linear relationships between predictors and outcomes, and relax the proportional hazard assumption (Jiang, 2019; Taylor, 2011). Given the tree-based estimators in step 2, the optimal estimator in step 3 can be selected via cross-validation by tuning the number of basis functions from step 1 and tree-based parameters from step 2.

The remainder of this paper is organized as follows. In Section 2, we define notation and describe the model setup. In particular, we give detailed discussion on the multivariate principal component analysis (MFPCA) for feature extraction from multiple time-varying covariates, and the construction of the functional ensemble survival tree for conducting individualized dynamic prediction. We investigate the finite sample performance in intensive simulation studies in Section 3, and provide

publicly available code in our R package funest (Xie et al., 2020). We illustrate the propose method in the ADNI dataset in Section 4, and conclude the paper with remarks and future research in Section 5.

2 | NOTATION AND METHOD

2.1 | Functional ensemble survival tree

Random survival forest (RSF) is an ensemble tree method that has been widely adopted for the analysis of right-censored survival data. The goal of constructing the RSF is to train a model that learns from the available functional and baseline covariates in the cohort, such that the model can be used to make risk predictions for new patients conditioning on partially observed data. The focus of this subsection is on model construction and we elaborate on individualized dynamic prediction in Section 2.2.

Typical RSF cannot take longitudinal covariates directly as inputs. To extend the survival tree on the basis of longitudinal covariates, we first characterize the changing patterns of the time-varying biomarkers via MFPCA. We start by setting up the functional framework for single time-varying biomarkers and then expand to the multivariate setting. We let $Y_i = (Y_{i,1}, \ldots, Y_{i,Q})$ be the observed time-varying biomarkers for individual i, $i = 1, \ldots, n$. The qth time-varying biomarker is denoted by $Y_{i,q} = (Y_{i,q}(t_{i,1}), \ldots, Y_{i,q}(t_{i,R_i}))'$ where R_i reflects random and irregular individual-specific visits, $q = 1, \ldots, Q$. We assume that the qth observed trajectory, $\forall\, q \in \{1, \ldots, Q\}$, is recorded with error,

$$Y_{i,q}(t_{i,r}) = Z_{i,q}(t_{i,r}) + \epsilon_{i,q,r}\,, \quad \forall t_{i,r} \in [0, \tau] \tag{1}$$

where $Z_{i,q}(t_{i,r})$ denotes the denoised mean value of $Y_{i,q}(t_{i,r})$ for $t_{i,r} \in [0, \tau]$ and τ denotes the maximum follow-up time in the cohort. The error term is assumed to have $E(\epsilon_{i,q,r}) = 0$ and $var(\epsilon_{i,q,r}) = \sigma_q^2$ (Yao et al., 2005).

Under the functional framework, we assume that $Z_{i,q} = \{Z_{i,q}(t), \forall\, t \in [0, \tau]\}$ are realizations of a stochastic process $Z_q(t)$ in a square integrable functional space with domain $[0, \tau]$. The stochastic process is assumed to have mean function $E[Z_q(t)] = \mu_q(t)$ and covariance operator $C_q(t, s) = Cov(Z_q(t), Z_q(s))$ for $\forall\, t, s \in [0, \tau]$. Then by Mercer's theorem (Mercer, 1909),

$$C_q(t, s) = \sum_{j=1}^{\infty} \lambda_j^q \phi_j^q(s) \phi_j^q(t),\ \forall\, t,\ s \in [0,\ \tau], \tag{2}$$

where $\phi_j^q(t)$ is the jth orthonormal eigenfunction and λ_j^q is the corresponding eigenvalue where $\lambda_1^q \geq \lambda_2^q, \geq \ldots, > 0$, $j = 1, \ldots, \infty$. This decomposition thus allows us to characterize each functional observation $Z_{i,q}(t)$ as

$$Z_{i,q}(t) = \mu_q(t) + \sum_{j=1}^{\infty} \xi_{i,j}^q \phi_j^q(t),\ \forall t \in [0,\ \tau], \tag{3}$$

where $\xi_{i,j}^q = \langle Z_{i,q} - \mu_q,\ \phi_j^q \rangle = \int_{t=0}^{\tau} [Z_{i,q}(t) - \mu_q(t)] \phi_j^q(t) dt$, is the jth functional principal component (FPC) score for individual i. According to the Karhunen–Loève theorem (Ramsay & Silverman, 2004), each curve $Z_{i,q}(t)$, $\forall t \in [0, \tau]$, can then be characterized by the infinite sequence of FPC scores $\xi_{i,j}^q$, $j = 1, \ldots, \infty$. In practice, an approximation of Equation (3) is usually carried out by truncating the infinite summation to the first M_q terms where M_q could be determined by, for example, Akaike

information criterion (AIC) or the total variance explained (TVE) (Wang et al., 2016). For estimation, given the observed data, we adopt the Principal Analysis by Conditional Estimation (PACE) algorithm for its well-known property of accommodating sparse longitudinal observations as is the case in our motivating study (Yao et al., 2005). Specifically, we use the PACE algorithm to facilitate the estimation of the discretized mean function $\widehat{\mu}_{i,q} = (\widehat{\mu}_q(t_{i,1}), \ldots, \widehat{\mu}_q(t_{i,R_i}))'$, the $R_i \times R_i$ empirical covariance matrix $\widehat{\Sigma}_i^q$, and the corresponding eigenvectors $\widehat{\phi}_j^q$ and eigenvalues $\widehat{\lambda}_j^q$, $j = 1, \ldots, M_q$, (Yao et al., 2005). Then the univariate FPC scores for the qth biomarker trajectory for ith individual can be estimated as

$$\widehat{\xi}_{i,j}^q = E(\xi_{i,j}^q | Y_{i,q}) = \widehat{\lambda}_j^q (\widehat{\phi}_j^q)^T (\widehat{\Sigma}_i^q)^{-1} (Y_{i,q} - \widehat{\mu}_{i,q}), \tag{4}$$

$j = 1, \ldots, M_q$, for $q = 1 \ldots, Q$.

Next we combine the Q univariate time-varying biomarkers via MFPCA following Happ and Greven (2018). We let $M = \sum_{q=1}^{Q} M_q$ and $\widehat{\Lambda} \in \mathbb{R}^{n \times M}$ be an $n \times M$ matrix for which the ith row is $\{\widehat{\xi}_{i,1}^1, \ldots, \widehat{\xi}_{i,M_1}^1, \ldots, \widehat{\xi}_{i,1}^Q, \ldots, \widehat{\xi}_{i,M_Q}^Q\}$. In the multivariate setting we aim to perform a matrix eigenanalysis such that we can estimate the corresponding eigenvectors $\widehat{v}_m$, from the empirical block matrix $\widehat{G} = \frac{1}{n-1} \widehat{\Lambda}^T \widehat{\Lambda} \in \mathbb{R}^{M \times M}$, $m = 1, \ldots, M$. Note that MFPCA indirectly accommodates the potential correlations among multiple trajectories via correlation among the FPC scores by pooling all estimated eigenvectors from the univariate biomarkers in the block matrix $\widehat{G}$. The eigenvectors $\widehat{v}_m$ thus contain the information of correlations across different time-varying biomarkers. As a result, the multivariate eigenfunctions are estimated as

$$\widehat{\psi}_m^q(t_q) = \sum_{k=1}^{M_q} [\widehat{v}_m]_k^q \widehat{\phi}_k^q(t_q), \quad t_q \in [0, \tau], \tag{5}$$

where $[\widehat{v}_m]_k^q$ denotes the kth entry in the qth block of $\widehat{v}_m$, $q = 1, \ldots, Q$, $m = 1, \ldots, M$. The corresponding individual-specific MFPC scores can thus be estimated as

$$\widehat{\rho}_{i,m} = \sum_{q=1}^{Q} \sum_{k=1}^{M_q} [\widehat{v}_m]_k^q \widehat{\xi}_{i,k}^q, \tag{6}$$

$m = 1, \ldots, M$. Similar to the univariate setting, the optimal number of MFPCs, $\{D: D \leq M\}$, can be chosen based on, for example, TVE or AIC.

The RSF can be easily constructed once the MFPCA scores have been estimated. Growing a single tree is well known to exhibit high variances in the predicted outcomes, combining a number of trees in forming a forest thus can substantially decrease the variability in prediction. Each tree within the forest is non-deterministic, meaning that is it grown on a subspace of individuals from bootstrapping the whole dataset (Ishwaran et al., 2008, 2011). Each tree is grown from a single node to a tree with multiple terminal nodes (i.e. nodes with no further split). Specifically, a tree is grown by partitioning individuals at each node into two groups, where the split is chosen under a user-specified splitting rule. Node splitting rules often are determined with the goal to either maximize within-node homogeneity or between-node heterogeneity. The standard split criterion for survival trees is the log-rank statistic to compare the survival distributions at each node which has been widely used and implemented (Ishwaran et al., 2008, 2011). Other splitting criterion such as the maximally selected rank statistic has been recently developed for its well-known unbiased split variable selection property (Wright et al., 2017). Each tree within the forest will undergo repeated binary splitting until a stopping criterion is

met. Lastly, the information from all trees is aggregated from the terminal nodes for an average risk prediction ensemble.

2.2 | Individualized dynamic prediction

We let n denote the number of individuals in the training cohort and $n + 1$ be the new individual who is event free and has observation up to some time t^*, $t^* < \tau$. For each single tree b, $b = 1, \ldots, B$, prediction of the survival probability at $t^* + \Delta t < \tau$, is made by dropping the new individual $n + 1$'s observations down the tree as

$$\widehat{S}_b(t^* + \Delta t | t^*) = \frac{\widehat{S}_b(t^* + \Delta t | W_{n+1}, \widehat{\rho}_{n+1})}{\widehat{S}_b(t^* | W_{n+1}, \widehat{\rho}_{n+1})}, \tag{7}$$

where W_{n+1} is the baseline covariates for individual $n + 1$ of dimension $P \times 1$. The MFPC scores $\widehat{\rho}_{n+1}$ can be obtained by first estimating the univariate FPC scores from (4),

$$\widehat{\xi}^q_{n+1,j} = \widehat{\lambda}^q_j (\widehat{\phi}^q_j)^T (\widehat{\Sigma}^q_{n+1})^{-1} (Y_{n+1,q} - \widehat{\mu}_{n+1,q}), \tag{8}$$

$j = 1, \ldots, M_q$, $q = 1, \ldots, Q$. We then pass these FPC scores to (6) to obtain the MFPCA scores $\widehat{\rho}_{n+1} = (\widehat{\rho}_{n+1,1}, \ldots, \widehat{\rho}_{n+1,D})'$. The final prediction from the forest is estimated by averaging over B trees,

$$\widehat{S}(t^* + \Delta t | t^*) = \frac{1}{B} \sum_{b=1}^{B} \widehat{S}_b(t^* + \Delta t | t^*). \tag{9}$$

3 | SIMULATION STUDY

We conduct intensive simulation studies to investigate the finite sample performance of our proposed method in this section. We aim to mimic the motivating application and simulate $n = 400$ individuals in each dataset, with $nsim = 500$ datasets. The individual-specific visit times $\{t_{i,r}, r = 1, 2, \ldots, 7\}$ are generated from the Gaussian distribution centred at 0, 3, 6, 9, 12, 15, and 18 with standard deviation of 0.1 except the initial baseline visit which is fixed at 0.

We assume that the time-varying biomarkers are recorded with error, $Y_{i,q}(t_{i,r}) = Z_{i,q}(t_{i,r}) + \epsilon_{i,r,q}$, where $\epsilon_{i,r,q} \sim N(0, 1)$ and $q = 1, 2, 3$. We consider both the linear and non-linear longitudinal trajectories in a similar fashion as Li and Luo (2019). Specifically in the linear setting, we simulate

$$Z_{i,q}(t_{i,r}) = \beta_{0q} + \beta_{tq} t_{i,r} + \beta_{1q} X_{i,q} + b_{i,q},$$

where $[\beta_{01}, \beta_{02}, \beta_{03}] = [1.5, 2, 0.5]$, $[\beta_{t1}, \beta_{t2}, \beta_{t3}] = [1.5, -1, 0.6]$, and $[\beta_{11}, \beta_{12}, \beta_{13}] = [2, -1, 1]$. We simulate $X_{i,q} \sim N(3, 1)$ for $q = 1, 2, 3$, and the individual-specific random effects $[b_{i,1}, b_{i,2}, b_{i,3}] \sim MVN(\mathbf{0}, \Sigma)$ with

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \eta_{12}\sigma_1\sigma_2 & \eta_{13}\sigma_1\sigma_3 \\ & \sigma_2^2 & \eta_{23}\sigma_2\sigma_3 \\ & & \sigma_3^2 \end{bmatrix},$$

where $[\sigma_1^2, \sigma_2^2, \sigma_3^2] = [1, 1.5, 2]$, and $[\eta_{12}, \eta_{13}, \eta_{23}] = [-0.2, 0.1, -0.3]$.

The non-linear trajectories for each individual i is assumed to follow a piecewise model

$$Z_{iq}(t_{ir}) = \beta_{0q} + \beta_{tq} \sum_{r=1}^{3} c_r (t_{ir} - k_r)^{(+)} + \beta_{1q} X_{iq} + b_{iq},$$

where the spline coefficients $[c_1, c_2, c_3] = [1.2, 0.7, 0.5]$, knots $[k_1, k_2, k_3] = [0, 6, 13]$, and truncated basis functions of time

$$(t_{ir} - k_r)^{(+)} = \begin{cases} t_{ir} - k_r, & t_{ir} \geq k_r \\ 0, & \text{otherwise} \end{cases}$$

We assume a proportional hazards model in this simulation where the hazard function follows,

$$h_i(t) = h_0(t) \exp\{\sum_{p=1}^{P} \gamma_p W_{i,p} + \sum_{q=1}^{3} \alpha_q Z_{i,q}(t)\},$$

where $h_0(t) = \exp(-7)$, α_q is set to be $(0.1, -0.1, 0.2)$ for $q = 1, 2, 3$ respectively. We consider four different scenarios for the set of fixed covariates $W_i = (W_{i,1}, \ldots, W_{i,P})'$. In the first two scenarios we set $\rho = 0.2$ and 0.5 for $P = 20, 100$, respectively, to represent strong autoregressive dependence where $W_i \sim \text{MVN}(0, \Sigma^{(W)})$ with the (k, l)th component of $\Sigma^{(W)}$ define as,

$$\Sigma_{k,l}^{(W)} = \begin{cases} 1, & k = l \\ \rho^{|k-l|}, & i \neq j \end{cases}.$$

In the last two scenarios, we consider binary covariates with $P(W_{i,p} = 1) = 0.5$ similarly under $\rho = 0.2$ and 0.5 for $P = 20, 100$. We set the associated coefficients $\gamma_p = (-2.5, -0.5, -0.15, -0.15, -0.1)$ for $p = 1, \ldots, 5$, so that high values of $\gamma_1, \ldots, \gamma_5$ are associated with shorter times to the event, and $\gamma_p = 0$ for $p = 6, \ldots, P$. The elements of W_i with non-zero coefficients were chosen to give both weak and strong dependence within the set of important covariates accompanied with set of noise variables. With the setup above, we are then ready to simulate the survival time T_i, which can be generated from the inverse of the cumulative hazard function $H_i^{-1}(u|\mathscr{D}_i; \gamma, \alpha)$, where $\mathscr{D}_i$ is the observed individual-level data, $\gamma = (\gamma_1, \ldots, \gamma_P)$, $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, and $u \sim \text{unif}(0, 1)$ (Austin, 2012). We have simulated under the independent censoring scheme, where the censoring time is set to follow a uniform distribution $unif(0, C_m)$, and C_m is set such that the % of being censored by the end of the study is 30%.

Given the simulation settings for the time-varying/fixed covariates along with the survival times as the response variable, we now describe how the dynamic predictions are conducted in this simulation study. In evaluating the model prediction performance, we adopted both the AUC for measuring model discrimination (Li et al., 2015), and Brier score as a combined measure of discrimination and calibration (Schoop et al., 2008). To avoid overfitting, we employed a fivefold inner and outer cross-validation in a similar fashion as Wright et al. (2017). Specifically for the inner cross-validation, an optimal ensemble survival tree model was built and selected based on the best prediction performance by tuning the parameters in each fold. For each fold in the outer cross-validation, the prediction accuracy measure is recorded dynamically for each time window $(t^*, t^* + \Delta t]$ conditional on data

observed up to t^*, $t^* = 6, 9$, forecasting $t^* + \triangle t$ for $\triangle t = 3, 6$. Note that for individuals in the testing set, their trajectories beyond t^* need to be removed prior to their estimation of the MFPCs, much like what we would have in a real data setting.

Table 1 illustrates simulation results from the non-linear setting. Additional simulation results under the linear setting are provided within the Supplemental Material for interested readers. As shown in Table 1, the AUC outputted from the proposed method are in good agreement with the true AUC on average, confirming a satisfactory model discrimination. The Brier scores, on the other hand, are very close to zero on average, which confirms good model calibration. In addition, we see that the proposed model retains robust prediction performance when the signal-to-noise ratio (S:N) decreases from 5 : 15 to 5 : 95. It should be noted that the RSF does not delete or select any of the variables. All of the variables can influence the prediction, if that is indicated by the training data.

The proposed method has been implemented in our `funest` package (Xie et al., 2020) which utilizes the well developed `ranger` (Wright & Ziegler, 2017) that wraps the implementation of random forest in `C++`. The computational speed of our package is outstanding, as it takes less than 30 s for growing and making dynamic predictions on the functional ensemble survival forest with ∼2500 trees on a desktop with i7-7700 CPU. In addition, the package naturally takes advantage of the multi-core processor when running in a larger scale computational environment which warrants an even more promising computational speed.

4 | ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

The data used in this section are obtained from the Alzheimer's ADNI database†. We treat the conversion from MCI to AD as the time-to-event outcome and focus on 302 patients who have been

TABLE 1 Estimated mean AUC(t^*, $t^* + \triangle t$) and Brier score(t^*, $t^* + \triangle t$) under the non-linear setting via functional ensemble survival tree; $n = 400$, $nsim = 500$, S:N = signal-to-noise ratio

W_i	P	S:N	t^*	Δt	True AUC	AUC	BS
Normal	20	5:15	6	3	0.892	0.847	0.134
				6	0.904	0.864	0.147
			9	3	0.877	0.815	0.151
				6	0.896	0.830	0.147
	100	5:95	6	3	0.898	0.855	0.137
				6	0.913	0.870	0.142
			9	3	0.881	0.813	0.164
				6	0.899	0.827	0.157
Binary	20	5:15	6	3	0.827	0.781	0.085
				6	0.848	0.816	0.143
			9	3	0.843	0.805	0.123
				6	0.868	0.838	0.149
	100	5:95	6	3	0.836	0.792	0.083
				6	0.867	0.834	0.127
			9	3	0.854	0.819	0.114
				6	0.883	0.852	0.133

diagnosed with MCI in ADNI-1. Out of those who were diagnosed with MCI, 137 of them progressed to AD before the end of the study. Patients were assessed at baseline, 6, 12, 18, 24, 30 and 36 months in ADNI-1 with additional annual follow-ups in ADNI-2 resulting in an average follow-up period of 31.3 (sd = 12.2) months. The corresponding average number of visits recorded was 6.2 (sd = 2.0). Table 2 in the Supplemental Material shows the list of variables that we consider in this section. We focus on five time-varying neurocognitive markers as well as other baseline covariates that have been well studied in the Alzheimer's literature (Gomar et al., 2014; LaFerla et al., 2007; Mattson, 2004). All time-varying markers have been re-scaled to have a variance equal to 1 in this analysis, ensuring that larger scales of marker variability do not dominate the MFPCA characterization in stage 1. Specifically, for each time-varying neurocognitive markers Y_q, $q = 1, \ldots, 5$, we estimate a weight as,

$$w_q = \left(\int_0^\tau \widehat{C}_q(t,t)dt\right)^{-1} = \left(\int_0^\tau \widehat{var}(Z_q(t))dt\right)^{-1}. \tag{10}$$

Using these weights, the time-varying markers can then be re-scaled to $w_q^{1/2} Y_q$, with a variance of 1 (Happ & Greven, 2018).

We compared the proposed model, which we refer to as the full model, with the Cox proportional hazards model (MFPCCox) (Li & Luo, 2019). There did not appear to be any extreme non-linearity or outliers in the Schoenfeld residuals plot for each of the baseline covariates, thus the proportional hazards assumption was deemed reasonable. To avoid overfitting, we employed a fivefold inner and 10-fold outer cross-validation in this analysis. For the five univariate expansions in MFPCA, we have selected the number of FPCs that can explain ≥95% of the univariate variance. We selected three MFPCs that explained 88% of variation on average across the 10 training datasets, with the first MFPC explaining 67% of the total variation on average. Additionally, we compared the full model and MFPCCox with a baseline model (only includes baseline measures, including the baseline measures for the time-varying markers).

Figure 2 shows the average prediction accuracy recorded dynamically for each time window $(t^*, t^* + \Delta t]$ conditional on data observed up to t^*, $t^* = 6, 12, 18, 24$ (month), forecasting $t^* + \Delta t$ for $\Delta t = 6$ month. It is apparent that the full model achieved a better AUC(t^*, $t^* + \Delta t$) and BS(t^*, $t^* + \Delta t$) dynamically over all time points, compared to the baseline model on average. This suggests that the inclusion of the extracted features of time-varying covariates indeed facilitates a better model discrimination and calibration. The MFPCCox model, on the other hand, achieved similar prediction performance compared with the full model. To further demonstrate the advantage of incorporating multiple time-varying markers, we have considered five reduced models. Each reduced models is constructed with all available baseline covariates, along with one of the five time-varying neurocognitive markers. The results of the corresponding AUC(t^*, $t^* + \Delta t$) and BS(t^*, $t^* + \Delta t$) for these reduced models are included in the Supplemental Material for interested readers.

We further illustrate the variable importance ranking via the variable permutation importance measure as shown in Figure 3. A variable is identified as important if it exerts a positive effect on the prediction performance. A greater value of permutation measure on a variable implies that the variable is more important for the overall predictive accuracy; see Nembrini et al. (2018) for more details. As a result, we see from Figure 3 that the first principal component (PC1), on average, stands out with large permutation importance measure in relation to other variables. This finding suggests that the contribution of the time-varying covariates in predicting the progression of AD for those who are diagnosed as MCI is much greater relative to the fixed baseline covariates. Such finding is indeed in agreement with what we observe in Figure 2. In addition, note that the neurocognitive markers along with PACC (preclinical Alzheimer cognitive composite score) are top ranked in Figure 3, which conveys that the cognitive measures for MCI patients are the most predictive variables for their disease

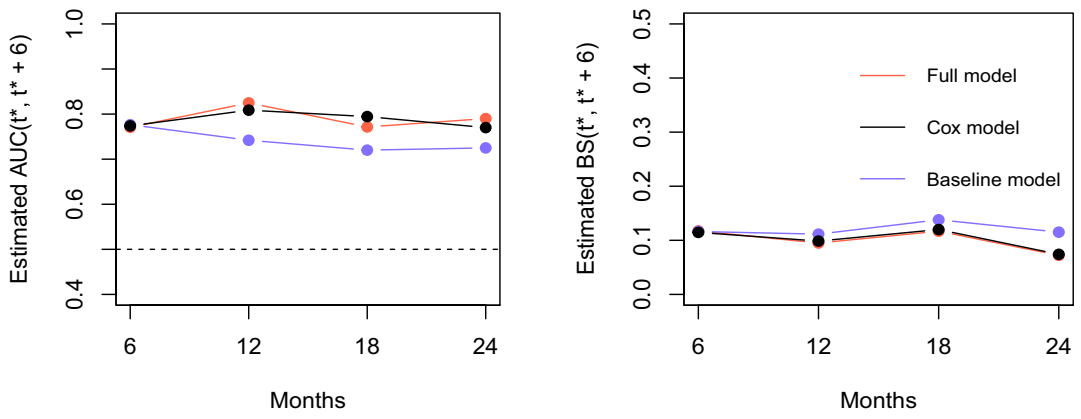

FIGURE 2 Comparison of dynamic prediction performances of the full, baseline and MFPCCox model averaged over cross-validations. AUC(t^*, $t^* + 6$) and BS(t^*, $t^* + 6$) conditional on data observed prior to $t^* = 6, 12, 18, 24$ (month) in forecasting $t^* + 6$ under a sliding window framework [Colour figure can be viewed at wileyonlinelibrary.com]

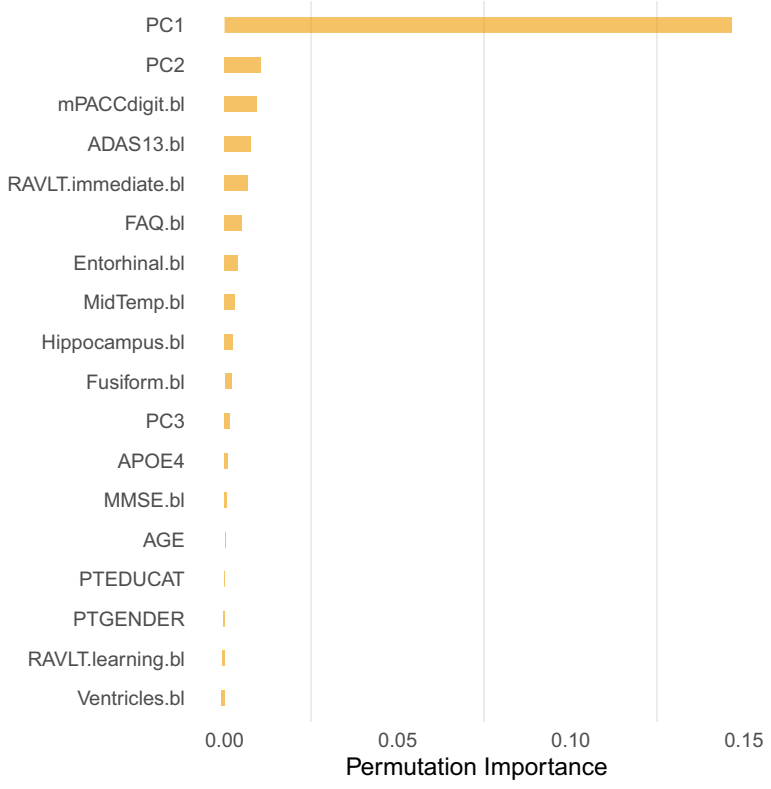

FIGURE 3 Variable permutation importance barplot averaged over cross-validations [Colour figure can be viewed at wileyonlinelibrary.com]

progression to AD. This finding is in accordance with the literature, as cognitive composite scores are known to retain great discriminative ability of the disease status within patients (Donohue et al., 2014; Rabin et al., 2019; Weintraub et al., 2018).

Lastly, we demonstrate individualized dynamic prediction where we randomly set aside a single patient from the cohort. The model was trained on the remainder of the cohort leaving this single patient out, such that we are able to visualize the predicted future biomarker trajectory and risk conditional on partial profile. Figure 4 shows two of the five neurocognitive markers (i.e. ADAS-COG13 and FAQ) that we have used in the model for ease of visualization. The vertical dashed line in Figure 4 represents the last time the biomarker has been recorded for the patient. From the first column of Figure 4, we can see that the predicted trajectories of the ADAS-COG13 and FAQ are in great harmony with the true values (blue and black asterisks, respectively). Correspondingly, the predicted AD-free probability for the patient is shown as a function of time in the second column where splines have been adopted to provide a smooth curve. The predicted low risk should be consistent with what one would expect from the stability of neurocognitive marker measurements. A full set of neurocognitive markers and their associated predictions are provided in the Supplemental Material for interested readers. The `R` package `funest` can be readily adopted for conducting the analysis in this section. We also provide sample code through an online repository (https://github.com/jj113/funest).

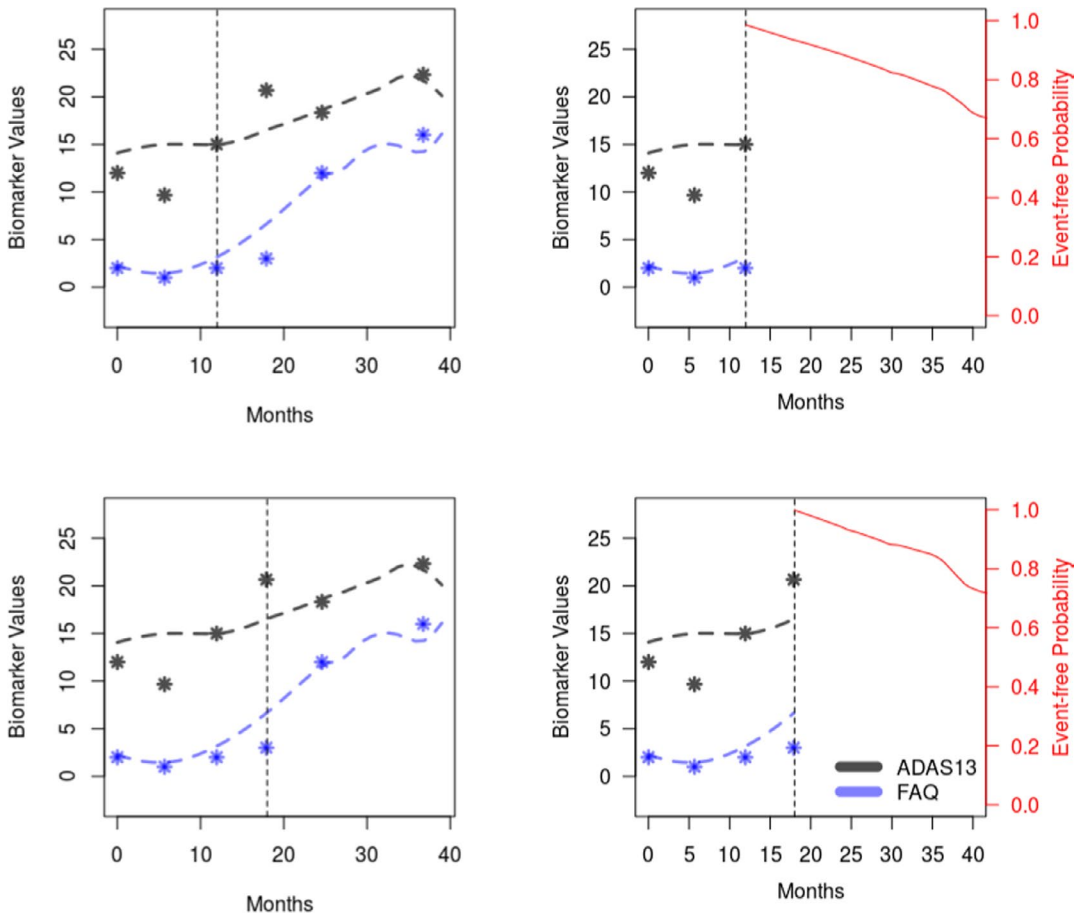

FIGURE 4 Predicted trajectories of ADAS-COG13 and FAQ in the first column and predicted AD-free probability in the second column conditional on partially observed marker values prior to the vertical dashed line; vertical dashed line represents the last time the biomarker has been recorded for the patient [Colour figure can be viewed at wileyonlinelibrary.com]

5 | DISCUSSION

We have formulated the functional ensemble survival tree framework that facilitates individualized dynamic prediction for disease progression accommodating multiple time-varying biomarkers. The proposed framework is fully data-driven and therefore removes the burden of the need to impose model assumptions on both the time-varying trajectories and survival distribution. Specifically, we adopt MFPCA to characterize the changing pattern of the multivariate time-varying biomarkers which effectively captures the correlation among them. We then nest these extracted features into a non-parametric ensemble survival tree which accommodates dynamic prediction under the presence of high dimensionality of baseline covariates. We have investigated the empirical performance of the proposed algorithm and have shown satisfactory model discrimination and calibration. We conducted individualized dynamic predictions and illustrate the utility of the proposed framework in the ADNI dataset. This could help physicians to predict the future course of the biomarker trajectories as well as the associated risk of AD which in turn, could facilitate identifying high risk individuals for prevention trials and treatment interventions.

Under the two-stage framework, the functional scores from stage 1 are used as inputs to stage 2 in conducting the dynamic predictions, neglecting the estimation errors made in the first stage. A model check should be applied in stage 1 prior to carrying the scores over to stage 2, to prevent utilizing poorly estimated parameters. An *ad hoc* approach is to visualize the bias between the estimated individualized trajectories and their observed true values. This has been done in the ADNI application, as shown in the first column of Figure 4. In addition, the asymptotic properties of the functional approach has been well derived in the literature, enabling the estimation of the uncertainty from stage 1 (Happ & Greven, 2018; Yao et al., 2005). Alternative models may need to be fitted in stage 1 if the functional scores are associated with large uncertainty. In the prediction context, when comparing different models, the best model is the one with the smallest prediction error or, equivalently, the best fit (Houwelingen & Putter, 2011). Further simulation studies may facilitate the investigation of this issue.

In this paper, we have characterized the time-varying marker trajectories on a common time domain $[0, \tau]$ for all individuals. In the general survival context, individuals are constrained in having different length of longitudinal trajectories (predictor domains) upon the occurrence of event or censoring. Thus, utilizing a common time domain for all individuals may not be optimal. Methods that can facilitate trajectory characterization, accommodating various predictor domains, are greatly needed if the between-subject variability in the width of the time domain varies substantially or when the time domain is informative. While some previous work have been done in this area when looking at the data retrospectively, no work has been done in the prediction context (Gellar et al., 2014; Johns et al., 2019). This will remain as part of our future work.

A limitation in all tree-based methods is the lack of interpretability. However, in analysing the ADNI dataset, we provided the variable permutation importance measure which identifies variables that are important contributors for the overall predictive accuracy. The comparison between baseline and full model also facilitates investigation of the added value of the time-varying covariates. Our findings from ADNI suggest that time-varying trajectories play a major role in predicting AD progression for those that are diagnosed with MCI. ADNI is an ongoing project that currently only contain sparse number of individuals who have their genetic profiles available. The model setup and the software distributed in this article warrants further research incorporating a richer set of genetic markers.

ACKNOWLEDGEMENTS

This work is supported in part by funding from the Foundation for Barnes Jewish Hospital and P30 CA091842. The authors wish to thank the referees of this article for providing helpful comments

and suggestions. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how to apply/ADNIAcknowledgement List.pdf.

REFERENCES

Austin, P.C. (2012) Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29), 3946–3958.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., et al. (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, 56(2), 766–781.

Donohue, M.C., Sperling, R.A., Salmon, D.P., Rentz, D.M., Raman, R., Thomas, R. G., et al. (2014) The preclinical alzheimer cognitive composite: Measuring amyloid-related decline. *JAMA Neurology*, 71(8), 961–970.

Ewers, M., Walsh, C., Trojanowski, J.Q., Shaw, L.M., Petersen, R.C., Jack, C.R., et al. (2012) Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiology of Aging*, 33(7), 1203–1214.e2.

Gellar, J.E., Colantuoni, E., Needham, D.M. & Crainiceanu, C.M. (2014) Variable-domain functional regression for modeling ICU data. *Journal of the American Statistical Association*, 109(508), 1425–1439.

Gomar, J.J., Conejero-Goldberg, C., Davies, P. & Goldberg, T.E. (2014) Extension and refinement of the predictive value of different classes of markers in ADNI: Four-year follow-up data. *Alzheimer's & Dementia*, 10(6), 704–712.

Happ, C. & Greven, S. (2018) Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522), 649–659.

Houwelingen, H.V. & Putter, H. (2011) *Dynamic prediction in clinical survival analysis*. Boca Raton: CRC Press

Ishwaran, H., Kogalur, U.B., Blackstone, E.H. & Lauer, M.S. (2008) Random survival forests. *The Annals of Applied Statistics*, 2(3), 841–860.

Ishwaran, H., Kogalur, U.B., Chen, X. & Minn, A.J. (2011) Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(1), 115–132.

Jiang, S. (2019) Prediction based on random survival forest. *American Journal of Biomedical Science & Research*, 6(2), 109–111.

Johns, J.T., Crainiceanu, C., Zipunnikov, V. & Gellar, J. (2019) Variable-domain functional principal component analysis. *Journal of Computational and Graphical Statistics*, 28(4), 993–1006.

Kong, D., Giovanello, K.S., Wang, Y., Lin, W., Lee, E., Fan, Y., et al. (2015) Predicting alzheimer's disease using combined imaging-whole genome snp data. *Journal of Alzheimer's Disease*, 46(3), 695–702.

Kong, D., Ibrahim, J.G., Lee, E. & Zhu, H. (2018) FLCRM: Functional linear Cox regression model. *Biometrics*, 74(1), 109–117.

LaFerla, F.M., Green, K.N. & Oddo, S. (2007) Intracellular amyloid-β in Alzheimer's disease. *Nature Reviews Neuroscience*, 8(7), 499–509.

Li, K. & Luo, S. (2019) Dynamic prediction of Alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Statistics in Medicine*, 38(24), 4804–4818.

Li, L., Hu, B. & Greene, T. (2015) A simple method to estimate the time-dependent ROC curve under right censoring.

Lin, J., Li, K. & Luo, S. (2020) Functional survival forests for multivariate longitudinal outcomes: Dynamic prediction of alzheimer's disease progression. *Statistical Methods in Medical Research*, 0962280220941532.

Mattson, M.P. (2004) Pathways towards and away from Alzheimer's disease. *Nature*, 430(7000), 631–639.

Mercer, J. (1909) Xvi. Functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209(441-458, 415–446.

Nembrini, S., König, I.R., & Wright, M.N. (2018) The revival of the Gini importance? *Bioinformatics,* 34(21), 3711–3718.

Park, L.Q. Gross, A.L. McLaren, D.J. Pa, J. Johnson, J.K., Mitchell, M. & Manly, J.J. (2012) Confirmatory factor analysis of the ADNI neuropsychological battery. *Brain Imaging and Behavior*, 6(4), 528–539.

Rabin, J.S., Klein, H., Kirn, D.R., Schultz, A.P., Yang, H.S., Hampton, O., et al. (2019) Associations of physical activity and β-Amyloid with longitudinal cognition and neurodegeneration in clinically normal older adults. *JAMA Neurology*, 76(10), 1203–1210.

Ramsay, J.O. & Silverman, B.W. (2004) *Functional data analysis*. Berlin: Springer Series in Statistics.

Risacher, S.L. Saykin, A.J. Wes, J.D. Shen, L. Firpi, H.A. & McDonald, B.C. (2009) Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Current Alzheimer Research*, 6(4), 347–361.

Rizopoulos, D. (2012) *Joint models for longitudinal and time-to-event data*. New York: Chapman and Hall/CRC.

Schoop, R., Graf, E. & Schumacher, M. (2008) Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*, 64(2), 603–610.

Taylor, J.M. (2011) Random survival forests. *Journal of Thoracic Oncology*, 6(12), 1974–1975.

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016) Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1), 257–295.

Weintraub, S., Carrillo, M.C., Farias, S.T., Goldberg, T.E., Hendrix, J.A., Jaeger, J., et al. (2018) Measuring cognition and function in the preclinical stage of alzheimer's disease. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 4, 64–75.

Welsh, A.H., Lin, X. & Carroll, R.J. (2002) Marginal longitudinal nonparametric regression. *Journal of the American Statistical Association*, 97(458), 482–493.

Wright, M.N. & Ziegler, A. (2017) Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.

Wright, M.N., Dankowski, T. & Ziegler, A. (2017) Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in medicine*, 36(8), 1272–1284.

Wu, C.O. & Chiang, C.-T. (2000) Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, 10(2), 433–456.

Xie, Y., Jiang, S. & Colditz, G.A. (2020) funest: Functional Ensemble Survival Tree for Dynamic Prediction. R package version 0.0.1.3.

Yan, F., Lin, X., Li, R. & Huang, X. (2018) Functional principal components analysis on moving time windows of longitudinal data: Dynamic prediction of times to event. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4), 961–978.

Yan, F., Lin, X. & Huang, X. (2017) Dynamic prediction of disease progression for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene. *The Annals of Applied Statistics*, 11(3), 1649–1670.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470), 577–590.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

R package: `funest` (https://cran.r-project.org/web/packages/funest/index.html). Additional simulation results and application findings for this article are available online.

How to cite this article: Jiang S, Xie Y, Colditz GA. Functional ensemble survival tree: Dynamic prediction of Alzheimer's disease progression accommodating multiple time-varying covariates. *J R Stat Soc Series C*. 2021;70:66–79. https://doi.org/10.1111/rssc.12449